

Robust mislabel logistic regression without modeling mislabel probabilities

Hung Hung^{*1}, Zhi-Yu Jou¹, and Su-Yun Huang²

¹Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taiwan

²Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Logistic regression is among the most widely used statistical methods for linear discriminant analysis. In many applications, we only observe possibly mislabeled responses. Fitting a conventional logistic regression can then lead to biased estimation. One common resolution is to fit a mislabel logistic regression model, which takes into consideration of mislabeled responses. Another common method is to adopt a robust M -estimation by down-weighting suspected instances. In this work, we propose a new robust mislabel logistic regression based on γ -divergence. Our proposal possesses two advantageous features: (1) It does not need to model the mislabel probabilities. (2) The minimum γ -divergence estimation leads to a weighted estimating equation without the need to include any bias correction term, i.e., it is automatically bias-corrected. These features make the proposed γ -logistic regression more robust in model fitting and more intuitive for model interpretation through a simple weighting scheme. Our method is also easy to implement, and two types of algorithms are included. Simulation studies and the Pima data application are presented to demonstrate the performance of γ -logistic regression.

Key words: Classification; Logistic regression; Minimum divergence estimation; Mislabeled response; Robust M -estimation.

1 Introduction

Logistic regression is one of the most widely used statistical methods for linear discriminant analysis. Let Y_0 be a binary response with $\{0, 1\}$ values, and X be the p -dimensional random vector of explanatory variables. Logistic regression assumes $P(Y_0 = 1|X = x)$ to satisfy the conditional label probability model

$$\pi(x; \beta) = \frac{\exp(\beta^\top x)}{1 + \exp(\beta^\top x)}, \quad (1)$$

where $\beta = (\beta_1, \dots, \beta_p)^\top$, and let β_0 denote the true value of β in model (1). The MLE is known to be the most efficient estimator for β_0 when data are truly generated from model (1). However, in some situations we can only observe a contaminated label Y instead of the true status Y_0 . That is, Y is flipped from Y_0 according to the *mislabel probabilities*

$$\eta_0(x) = P(Y = 1|Y_0 = 0, X = x) \quad \text{and} \quad \eta_1(x) = P(Y = 0|Y_0 = 1, X = x). \quad (2)$$

The success probability of Y no longer follows model (1), but instead has the form

$$P(Y = 1|X = x) = \eta_0(x) \{1 - \pi(x; \beta)\} + \{1 - \eta_1(x)\} \pi(x; \beta). \quad (3)$$

Fitting label contaminated data $\{(Y_i, X_i)\}_{i=1}^n$ to the uncontaminated model (1) will produce a biased estimate of β_0 . To overcome the problem of mislabeling, some robustified logistic regression methods are developed based on (3) with different modelings for $\eta_j(x)$'s. Copas (1988) considered equal and constant mislabel probabilities, $\eta_0(x) = \eta_1(x) = \eta$, which we call the *constant-mislabel logistic regression*. For any given η , the estimating equation of β is $\frac{1}{n} \sum_{i=1}^n w_{\eta,i}(\beta) \{Y_i - \pi_\eta(X_i; \beta)\} X_i = 0$ with $\pi_\eta(x; \beta) = \eta \{1 - \pi(x; \beta)\} + (1 - \eta) \pi(x; \beta)$ and the weight function

$$w_{\eta,i}(\beta) = \frac{1 - 2\eta}{\{1 - \eta + \eta \exp(-\beta^\top X_i)\} \{1 - \eta + \eta \exp(\beta^\top X_i)\}}. \quad (4)$$

Another example is the *asymmetric-mislabel logistic regression* (Wainer, Bradlow and Wang, 2007; Komori *et al.*, 2016), which assumes $\eta_0(x) = \eta$ and $\eta_1(x) = 0$, i.e., mislabeling occurs

only in the 0-group. Hayashi (2012) extended the work of η -boost (Takenouchi and Eguchi, 2004) to propose a robustified boosting method for binary classification, which is equivalent to assuming the following mislabel probabilities

$$\eta_j(x) = \frac{2\xi_j}{(1 - \xi_0 - \xi_1) \left\{ \exp(\frac{1}{2}\beta^\top x) + \exp(-\frac{1}{2}\beta^\top x) \right\} + 2(\xi_0 + \xi_1)}, \quad j = 0, 1 \quad (5)$$

with extra parameters $\xi = (\xi_0, \xi_1)$. We call the corresponding model the ξ -*logistic* regression. Note that (5) attains its maximum value at the classification boundary $\beta^\top x = 0$. For any given ξ , the estimating equation of β is $\frac{1}{n} \sum_{i=1}^n w_{\xi,i}(\beta) \{Y_i - \pi_\xi(X_i; \beta)\} X_i = 0$, where $\pi_\xi(x; \beta) = \eta_0(x) \{1 - \pi(x; \beta)\} + \{1 - \eta_1(x)\} \pi(x; \beta)$ with $\eta_j(x)$ given in (5), and the weight

$$w_{\xi,i}(\beta) = \{1 - \eta_0(X_i) - \eta_1(X_i)\} \nu(X_i; \beta) + \eta'_0(X_i) \{1 - \pi(X_i; \beta)\} - \eta'_1(X_i) \pi(X_i; \beta) \quad (6)$$

with $\eta'_j(x) = \frac{\partial \eta_j(x)}{\partial (\beta^\top x)}$ and $\nu(x; \beta) = \pi(x; \beta) \{1 - \pi(x; \beta)\}$. Robustness of all the above-mentioned methods come from the underlying weight functions. For instance, in the constant-mislabel logistic regression, instances with larger values of $|\beta^\top X_i|$ get less weight $w_{\eta,i}(\beta)$ in the estimating equation.

The MLE for the above-mentioned robust logistic regression models, where mislabel probabilities $\eta_j(x)$'s are assumed to take a certain parametric form, is known to be sensitive to the misspecification of $\eta_j(x)$'s. Modeling mislabel probabilities, however, may not be straightforward. In applications, often we are mainly interested in is the true success probability $P(Y_0 = 1|X)$ instead of the nuisance parameters $\eta_j(x)$'s. There seems to be less necessary to build models for $\eta_j(x)$'s. The aim of this paper is to develop a robust mislabel logistic inference procedure that avoids modeling $\eta_j(x)$'s. The main idea is to replace the minimum Kullback-Leibler (KL) divergence estimation, which corresponds to the MLE, with the minimum γ -divergence estimation, which we call γ -*logistic* regression.

The paper is organized as follows. In Section 2, we review γ -divergence and use it to propose our robust γ -logistic regression, while its asymptotic properties and comparisons

with existing methods are discussed in Section 3. Simulation studies and the Pima data analysis are placed in Sections 4-5. The paper ends with a discussion in Section 6.

2 Method: γ -Logistic Regression

2.1 The minimum γ -divergence estimation and its robustness

Let g be the data generating distribution and f_θ be the model distribution indexed by the parameter θ , and let θ_0 denote the true parameter value of interest. The γ -divergence between g and f_θ is defined to be

$$D_\gamma(g, f_\theta) = \frac{1}{\gamma(\gamma+1)} \left\{ \|g\|_{\gamma+1} - \int \left(\frac{f_\theta}{\|f_\theta\|_{\gamma+1}} \right)^\gamma g \right\}, \quad (7)$$

where $\|f_\theta\|_{\gamma+1} = (\int f_\theta^{\gamma+1})^{\frac{1}{\gamma+1}}$. This divergence is introduced in Jones *et al.* (2001) with the name *density power divergence of type-zero*. The name γ -divergence is later introduced in Fujisawa and Eguchi (2008). In the limiting case, $\lim_{\gamma \rightarrow 0} D_\gamma(g, f_\theta) = \int \ln(\frac{g}{f_\theta})g$, which is the KL-divergence. The estimation criterion of minimum γ -divergence estimates θ_0 by

$$\operatorname{argmin}_\theta D_\gamma(g, f_\theta) = \operatorname{argmax}_\theta \int \left(\frac{f_\theta}{\|f_\theta\|_{\gamma+1}} \right)^\gamma g. \quad (8)$$

When g belongs to the parametric class $\{f_\theta : \theta \in \Theta\}$ with the parameter value θ_0 , the problem (8) is optimized at $\theta = \theta_0$. It ensures the consistency of the minimum γ -divergence estimation. In the presence of contamination, however, $g = cf_{\theta_0} + (1-c)h$ which is a mixture of the target distribution f_{θ_0} and certain contamination distribution h , where $1-c$ denotes the contamination proportion. With some calculations, it leads to

$$D_\gamma(g, f_\theta) = \left\{ c D_\gamma(f_{\theta_0}, f_\theta) + \frac{B_\gamma(c, h; \theta)}{\gamma(\gamma+1)} \right\} + \frac{\|cf_{\theta_0} + (1-c)h\|_{\gamma+1} - c\|f_{\theta_0}\|_{\gamma+1}}{\gamma(\gamma+1)} \quad (9)$$

with $B_\gamma(c, h; \theta) = (1-c) \int \left(\frac{f_\theta}{\|f_\theta\|_{\gamma+1}} \right)^\gamma h$. Ignoring terms not involving θ , minimizing (9) over θ is equivalent to minimizing

$$c D_\gamma(f_{\theta_0}, f_\theta) + \frac{B_\gamma(c, h; \theta)}{\gamma(\gamma+1)} \approx c D_\gamma(f_{\theta_0}, f_\theta), \quad (10)$$

where the approximation holds provided that, for some γ , the bias $B_\gamma(c, h; \theta)$ is negligibly small for θ in a neighborhood of θ_0 . The right hand side of (10) is minimized at $\theta = \theta_0$. That is, the minimization process is less affected by the mixing proportion c and the contamination h and, hence, we can estimate θ_0 well with negligibly small bias. See Fujisawa and Eguchi (2008) and Kanamori and Fujisawa (2015) for further discussions.

2.2 γ -Logistic Regression

The robust γ -divergence can be used to infer model (1) when the data are actually generated from (3). The reasons are discussed below.

Theorem 1. *The distribution of contaminated Y in (3) can be expressed as a mixture of the target distribution $P(Y_0 = y|X = x)$ and the mislabel-induced distribution $h(y|x)$,*

$$P(Y = y|X = x) = c(x) P(Y_0 = y|X = x) + \{1 - c(x)\} h(y|x),$$

where $h(y|x) = \left\{ \frac{\eta_0(x)}{\eta_0(x) + \eta_1(x)} \right\}^y \left\{ \frac{\eta_1(x)}{\eta_0(x) + \eta_1(x)} \right\}^{1-y}$ and $1 - c(x) = \eta_0(x) + \eta_1(x)$ is the conditional contamination proportion given $X = x$.

Theorem 1 sheds some light on the possibility of inferring the true success probability $P(Y_0 = y|X)$ from the contaminated data (Y, X) , since γ -divergence is able to ignore the influence from $h(y|x)$ as revealed in (10). Specifically, our robust γ -logistic adopts the conventional logistic regression model

$$f(y|x; \beta) = \{\pi(x; \beta)\}^y \{1 - \pi(x; \beta)\}^{1-y} \quad (11)$$

for Y_0 , while the observed Y is generated from (3), or equivalently, from the mixture

$$g(y|x) = c(x) f(y|x; \beta_0) + \{1 - c(x)\} h(y|x), \quad (12)$$

where $c(x)$ and $h(y|x)$ are defined in Theorem 1. By substituting the model distribution $f(y|x; \beta)$ and the data distribution $g(y|x)$ into (8) and taking expectation with respect to

X , γ -logistic estimates β_0 by

$$\operatorname{argmin}_{\beta} E_X \left[D_{\gamma} \left\{ g(\cdot|X), f(\cdot|X; \beta) \right\} \right] = \operatorname{argmax}_{\beta} E_{X,Y} \left\{ \left(\frac{f(Y|X; \beta)}{\|f(\cdot|X; \beta)\|_{\gamma+1}} \right)^{\gamma} \right\}, \quad (13)$$

where $\|f(\cdot|x; \beta)\|_{\gamma+1} = [\{\pi(x; \beta)\}^{\gamma+1} + \{1 - \pi(x; \beta)\}^{\gamma+1}]^{1/(\gamma+1)}$, and E_X and $E_{X,Y}$ denote the expectation with respect to X and (X, Y) , respectively. Recall that the validity of minimum γ -divergence estimation relies on the approximation (10), where the bias term $B_{\gamma}(c, h; \beta)$ plays a key role. From the expressions of $f(y|x; \beta)$ in (11) and $(c(x), h(y|x))$ in Theorem 1, we derive in Supplementary Materials that

$$\begin{aligned} B_{\gamma}\{c(x), h(\cdot|x); \beta\} &= \eta_0(x) \left\{ \pi(x; (\gamma+1)\beta) \right\}^{\frac{\gamma}{\gamma+1}} + \eta_1(x) \left\{ 1 - \pi(x; (\gamma+1)\beta) \right\}^{\frac{\gamma}{\gamma+1}} \\ &\rightarrow \eta_0(x) I(\beta^{\top} x > 0) + \eta_1(x) I(\beta^{\top} x \leq 0) \quad \text{as } \gamma \rightarrow \infty, \end{aligned} \quad (14)$$

where $I(\cdot)$ is an indicator function. It implies that the robustness of γ -logistic can be ensured for a large γ , provided that $E_X\{\eta_0(X)I(\beta^{\top} X > 0)\}$ and $E_X\{\eta_1(X)I(\beta^{\top} X \leq 0)\}$ at $\beta \approx \beta_0$ are negligible, under which $E_X[B_{\gamma}\{c(X), h(\cdot|X); \beta\}]$ can only have limited influence on (13). See Remarks 2-3 for further discussions about the robustness of γ -logistic.

In the sample level, the robust estimator $\hat{\beta}_{\gamma}$ is obtained via the empirical version of (13),

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{f(Y_i|X_i; \beta)}{\|f(\cdot|X_i; \beta)\|_{\gamma+1}} \right)^{\gamma} = \frac{1}{n} \sum_{i=1}^n \left(\frac{\exp\{Y_i(\gamma+1)\beta^{\top} X_i\}}{1 + \exp\{(\gamma+1)\beta^{\top} X_i\}} \right)^{\frac{\gamma}{\gamma+1}}. \quad (15)$$

Direct differentiation of (15) leads to the estimating equation $S_{\gamma}(\hat{\beta}_{\gamma}) = 0$, where

$$S_{\gamma}(\beta) = \frac{1}{n} \sum_{i=1}^n w_{\gamma,i}(\beta) \left\{ Y_i - \pi(X_i; (\gamma+1)\beta) \right\} X_i \quad (16)$$

with the weight function

$$w_{\gamma,i}(\beta) = \left(\frac{\exp\{Y_i(\gamma+1)\beta^{\top} X_i\}}{1 + \exp\{(\gamma+1)\beta^{\top} X_i\}} \right)^{\frac{\gamma}{\gamma+1}}. \quad (17)$$

From (16)-(17), the robustness of $\hat{\beta}_{\gamma}$ is clear, as $w_{\gamma,i}(\beta)$ down-weights instances with non-matched $(Y_i, \beta^{\top} X_i)$. Note that the robustness of γ -logistic is controlled by the value of γ .

When $\gamma = 0$, the estimating equation reduces to the non-robust estimating equation, $\frac{1}{n} \sum_{i=1}^n \{Y_i - \pi(X_i; \beta)\} X_i = 0$, for the conventional logistic regression. On the other hand, a large γ corresponds to a robust estimate of β_0 , but at the cost of being less efficient than MLE. See Remark 4 for a selection method of γ . See also Supplementary Materials for two types of algorithms for implementing γ -logistic regression.

Besides the parameter estimation, another important issue is to identify mislabeled subjects. Kanamori and Fujisawa (2015) developed a method to estimate the expected mixing proportion, $c = E_X\{c(X)\}$, based on the density power divergence. With this estimated c , they proposed to identify $100(1 - c)\%$ subjects with the smallest estimated values of $f(Y_i|X_i; \beta_0)$ as outliers. On the other hand, the weight $w_{\gamma,i}(\hat{\beta}_\gamma)$ from γ -logistic can provide a measure of label confidence. It motivates us to identify mislabeled subjects by searching for instances with small values of $w_{\gamma,i}(\hat{\beta}_\gamma)$. To have an objective evaluation criterion, we obtain the p-values of $w_{\gamma,i}(\hat{\beta}_\gamma)$'s by parametric bootstrap. Let $w_{\gamma,i}^{(b)}$ be the b -th bootstrapped version of $w_{\gamma,i}(\hat{\beta}_\gamma)$ by the null data $\{(\hat{Y}_i^{(b)}, X_i)\}_{i=1}^n$, where $\hat{Y}_i^{(b)}$ is generated from model (1) given $X = X_i$ and $\beta = \hat{\beta}_\gamma$. The p-value of $w_{\gamma,i}(\hat{\beta}_\gamma)$ is

$$PV_i = \frac{1}{b'} \sum_{b=1}^{b'} I \left\{ w_{\gamma,i}^{(b)} \leq w_{\gamma,i}(\hat{\beta}_\gamma) \right\} \quad (18)$$

for a large b' . Instances, e.g., $\{i : PV_i < 0.01\}$, can be identified for further examination.

We close this section by giving a few remarks on the robustness of γ -logistic, the confounding issue of the model misspecification and mislabeling, and the selection of γ value.

Remark 2 (robustness). *For symmetric mislabeling $\eta_0(x) = \eta_1(x)$, equation (14) becomes $\lim_{\gamma \rightarrow \infty} B_\gamma\{c(x), h(\cdot|x); \beta\} = \eta_0(x) = \eta_1(x)$, which does not involve the parameter β , i.e., the bias term $B_\gamma\{c(x), h(\cdot|x); \beta\}$ plays no role in parameter estimation as $\gamma \rightarrow \infty$. In other words, γ -logistic with a large γ produces a consistent estimate of β_0 regardless of the functional forms of $\eta_0(x)$ and $\eta_1(x)$ as long as $\eta_0(x) = \eta_1(x)$.*

Remark 3 (confounding). *In all previous discussions, we assume the model is correctly specified, i.e., $P(Y_0 = 1|X = x) = \pi(x; \beta_0)$ for some β_0 . When the model is misspecified, there is no so-called true β_0 , and the target parameter is criterion-dependent. With γ -divergence, the target parameter is $\beta_\gamma^* = \operatorname{argmin}_\beta E_X[D_\gamma\{f_{Y_0|X}(\cdot|X), f(\cdot|X; \beta)\}]$ with $f_{Y_0|X}(y|x) = P(Y_0 = y|X = x)$. Similar to the derivation of (9) with $g(y|x) = c(x)f_{Y_0|X}(y|x) + \{1 - c(x)\}h(y|x)$, we have (up to terms without involving β)*

$$D_\gamma\{g(\cdot|x), f(\cdot|x; \beta)\} \propto c(x) D_\gamma\{f_{Y_0|X}(\cdot|x), f(\cdot|x; \beta)\} + \frac{B_\gamma\{c(x), h(\cdot|x); \beta\}}{\gamma(\gamma + 1)}$$

with the bias term $B_\gamma\{c(x), h(\cdot|x); \beta\}$ being defined in (14). Note that $B_\gamma\{c(x), h(\cdot|x); \beta\}$ does not involve $f_{Y_0|X}(y|x)$, and the robustness of γ -logistic in estimating β_γ^ is still valid for a large γ , provided that $E_X\{\eta_0(X)I(\beta^\top X > 0)\}$ and $E_X\{\eta_1(X)I(\beta^\top X \leq 0)\}$ at $\beta \approx \beta_\gamma^*$ are small as discussed in texts below (14). Moreover, by Remark 2, the robustness of γ -logistic is unaffected by the functional forms of $\eta_0(x)$ and $\eta_1(x)$ when $\eta_0(x) = \eta_1(x)$.*

Remark 4 (selection of γ). *One can use the idea of Mollah, Eguchi and Minami (2007) to select γ by $\operatorname{argmax}_\gamma \frac{1}{n} \sum_{i=1}^n w_{\gamma_0, i}(\widehat{\beta}_\gamma)$ from (15), where γ_0 is a predetermined reference value, e.g., $\gamma_0 = 0.1$. Note that (15) can be affected by mislabeled Y_i . Thus, alternatively we replace the weight by its conditional expectation, i.e., $E[w_{\gamma_0, i}(\beta_0)|X_i] = \|f(\cdot|X_i; \beta_0)\|_{\gamma_0+1}$, and propose to select γ by $\operatorname{argmax}_\gamma \frac{1}{n} \sum_{i=1}^n \|f(\cdot|X_i; \widehat{\beta}_\gamma)\|_{\gamma_0+1}$.*

3 Characteristics of γ -Logistic Regression

3.1 Influence function and asymptotic properties of $\widehat{\beta}_\gamma$

Since $\widehat{\beta}_\gamma$ is an M -estimator, the influence function $\operatorname{IF}_{\widehat{\beta}_\gamma}(X_i, Y_i)$ of $\widehat{\beta}_\gamma$ evaluated at (Y_i, X_i) and $\beta = \beta_0$ is the negative Hessian inverse times the i -th element of the score function:

$$\operatorname{IF}_{\widehat{\beta}_\gamma}(X_i, Y_i) = w_{\gamma, i}(\beta_0) \left\{ Y_i - \pi(X_i; (\gamma + 1)\beta_0) \right\} H_\gamma^{-1} X_i, \quad (19)$$

where $H_\gamma = E[-\frac{\partial}{\partial \beta} S_\gamma(\beta)|_{\beta=\beta_0}] = E[w_{\gamma,i}(\beta_0) \nu(X_i; (\gamma+1)\beta_0) X_i X_i^\top] + \Delta_\gamma$ with

$$\Delta_\gamma = \gamma E \left[w_{\gamma,i}(\beta_0) \left[\nu(X_i; (\gamma+1)\beta_0) - \left\{ Y_i - \pi(X_i; (\gamma+1)\beta_0) \right\}^2 \right] X_i X_i^\top \right]$$

and $\nu(x; \beta) = \pi(x; \beta)\{1 - \pi(x; \beta)\}$. Direct calculation gives $\Delta_\gamma = 0$, and H_γ reduces to

$$H_\gamma = E \left[\|f(\cdot|X_i; \beta_0)\|_{\gamma+1} \nu(X_i; (\gamma+1)\beta_0) X_i X_i^\top \right]. \quad (20)$$

The robustness of γ -logistic can be seen from $\text{IF}_{\hat{\beta}_\gamma}(X_i, Y_i)$, where a large difference $\{Y_i - \pi(X_i; (\gamma+1)\beta_0)\}$ (which occurs when Y_i is mislabeled) will accompany with a small value of $w_{\gamma,i}(\beta_0)$, so that the influence of mislabeling is mitigated. As to the case of conventional logistic regression, which corresponds to $\text{IF}_{\hat{\beta}_\gamma}(X_i, Y_i)$ with $\gamma = 0$, we have $w_{\gamma,i}(\beta_0) = 1$ and there is no chance to achieve robustness when Y_i is mislabeled.

The asymptotic normality of γ -logistic is established below.

Theorem 5. *Assume the validity of model (1) and $E\|\text{IF}_{\hat{\beta}_\gamma}(X, Y)\|^2 < \infty$. As $n \rightarrow \infty$, we have the weak convergence $\sqrt{n}(\hat{\beta}_\gamma - \beta_0) \xrightarrow{d} N(0, \Sigma_\gamma)$, where $\Sigma_\gamma = H_\gamma^{-1} U_\gamma H_\gamma^{-1}$, H_γ is defined in (20), and $U_\gamma = E[w_{\gamma,i}^2(\beta_0) \{Y_i - \pi(X_i; (\gamma+1)\beta_0)\}^2 X_i X_i^\top]$.*

The asymptotic covariance matrix Σ_γ can be estimated by the sandwich-type estimator

$$\hat{\Sigma}_\gamma = \left\{ \hat{H}_\gamma(\hat{\beta}_\gamma) \right\}^{-1} \hat{U}_\gamma(\hat{\beta}_\gamma) \left\{ \hat{H}_\gamma(\hat{\beta}_\gamma) \right\}^{-1}, \quad (21)$$

where

$$\begin{aligned} \hat{U}_\gamma(\beta) &= \frac{1}{n} \sum_{i=1}^n w_{\gamma,i}^2(\beta) \{Y_i - \pi(X_i; (\gamma+1)\beta)\}^2 X_i X_i^\top \\ \hat{H}_\gamma(\beta) &= \frac{1}{n} \sum_{i=1}^n \|f(\cdot|X_i; \beta)\|_{\gamma+1} \nu(X_i; (\gamma+1)\beta) X_i X_i^\top + \hat{\Delta}_\gamma(\beta) \\ \hat{\Delta}_\gamma(\beta) &= \frac{\gamma}{n} \sum_{i=1}^n w_{\gamma,i}(\beta) \left[\nu(X_i; (\gamma+1)\beta) - \{Y_i - \pi(X_i; (\gamma+1)\beta)\}^2 \right] X_i X_i^\top. \end{aligned}$$

Note that we still include $\hat{\Delta}_\gamma(\hat{\beta}_\gamma)$ in $\hat{H}_\gamma(\hat{\beta}_\gamma)$ to estimate $\Delta_\gamma = 0$, since its effect cannot be ignored under finite samples. Subsequent inference about β_0 can be based on $(\hat{\beta}_\gamma, \hat{\Sigma}_\gamma)$.

3.2 Comparison with model-based mislabel logistic regression

A major difference between γ -logistic and the model-based mislabel logistic, e.g., constant-mislabel logistic and ξ -logistic, is the weight functions (see Figure 1). The weights $w_{\eta,i}(\beta)$ and $w_{\xi,i}(\beta)$ depend on $\beta^\top X_i$ only, which always down-weights samples with large $|\beta^\top X_i|$ values. Among these instances with large $|\beta^\top X_i|$ values, some are correctly-labeled. On the other hand, the weight $w_{\gamma,i}(\beta)$ of γ -logistic depends on both (Y_i, X_i) , and it only down-weights instances with non-matched $(Y_i, \beta^\top X_i)$. γ -logistic is able to weigh data instances in a more correct way, and thus can be expected to perform better than model-based mislabel logistic regressions under severe contamination. Another advantage is that the validity of γ -logistic mainly relies on putting less weight on instances having non-matched $(Y_i, \beta^\top X_i)$, and does not rely on any modeling of the mislabel probabilities $\eta_j(x)$'s. As for model-based mislabel logistic regressions, they incorporate the mislabel probabilities into model (3), which requires a further modeling for the nuisance parameters $\eta_j(x)$'s. The form of $\eta_j(x)$, however, is rarely known in practice, and the performance of model-based mislabel logistic can be questionable when complicated mislabel probabilities are present.

3.3 Comparison with robust mislabel logistic regression using density power divergence

Ghosh and Basu (2016) proposed a robust GLM by the minimum density power divergence estimation. For any $\alpha > 0$, the density power divergence between g and f_θ is

$$D_\alpha(g, f_\theta) = \alpha \int f_\theta^{\alpha+1} - (\alpha + 1) \int g f_\theta^\alpha + \int g^{\alpha+1}. \quad (22)$$

The estimating equation by replacing D_γ in (13) with D_α becomes $S_\alpha(\hat{\beta}_\alpha) = 0$, where

$$S_\alpha(\beta) = \frac{1}{n} \sum_{i=1}^n \left[w_{\alpha,i}(\beta) \left\{ Y_i - \pi(X_i; \beta) \right\} - b_\alpha(X_i; \beta) \right] X_i \quad (23)$$

with the weight $w_{\alpha,i}(\beta) = \left\{ \frac{\exp(Y_i \beta^\top X_i)}{1 + \exp(\beta^\top X_i)} \right\}^\alpha$ and $b_\alpha(x; \beta) = \frac{\exp(\beta^\top x) \{ \exp(\alpha \beta^\top x) - 1 \}}{\{1 + \exp(\beta^\top x)\}^{2+\alpha}}$ being the bias correction term. The following result is established by Ghosh and Basu (2016).

Theorem 6 (Ghosh and Basu, 2016). *Under model (1), the influence function of $\hat{\beta}_\alpha$ evaluated at (Y_i, X_i) and $\beta = \beta_0$ is $\text{IF}_{\hat{\beta}_\alpha}(X_i, Y_i) = [w_{\alpha,i}(\beta_0) \{Y_i - \pi(X_i; \beta_0)\} - b_\alpha(X_i; \beta_0)] H_\alpha^{-1} X_i$, where $H_\alpha = E[\xi_\alpha(X_i; \beta_0) \nu(X_i; \beta_0) X_i X_i^\top]$ with $\xi_\alpha(x; \beta) = \frac{\exp(\alpha \beta^\top x) + \exp(\beta^\top x)}{\{1 + \exp(\beta^\top x)\}^{1+\alpha}}$. Moreover, $\sqrt{n}(\hat{\beta}_\alpha - \beta_0) \xrightarrow{d} N(0, \Sigma_\alpha)$ with $\Sigma_\alpha = H_\alpha^{-1} U_\alpha H_\alpha^{-1}$ and $U_\alpha = E[\xi_\alpha^2(X_i; \beta_0) \nu(X_i; \beta_0) X_i X_i^\top]$.*

For simplicity in notation, we use the term α -logistic to denote the Ghosh-Basu logistic regression, since the density power divergence D_α is indexed by α . Although both γ -logistic and α -logistic are derived from the minimum divergence estimation, they have different behaviors in estimating β_0 . First, the robustness of both methods comes from the weight functions $w_{\gamma,i}(\beta)$ and $w_{\alpha,i}(\beta)$, and they are connected via $\{w_{\gamma,i}(\beta)\}^{\gamma+1} = w_{\alpha,i}((\gamma+1)\beta)$ when $\gamma = \alpha$. It indicates that the two methods share the same spirit of robustness. However, the resulting estimating equations are quite different in the bias correction scheme. In particular, γ -logistic corrects the bias by using the expanded parameter $(\gamma+1)\beta$ in (16), while α -logistic subtracts a bias correction term $b_\alpha(x; \beta)$ in (23). A consequence is that $S_\gamma(\beta)$ of γ -logistic consists of a weighted sum expression with the weight $w_{\gamma,i}(\beta)$, which directly reflects the contribution of the i -th instance to the estimator $\hat{\beta}_\gamma$, while this is not the case for $S_\alpha(\beta)$ of α -logistic. Another difference is the ability of robustness. As shown in (10), γ -divergence is able to ignore the influence of mislabeling, and we can expect a strong robustness property for γ -logistic. However, this is not the case for the density power divergence D_α . This can be seen from the fact that, when $g = cf_{\theta_0} + (1-c)h$, we have

$$D_\alpha(g, f_\theta) \propto D_\alpha(cf_{\theta_0}, f_\theta) - (1-c) \int f_\theta^\alpha \approx D_\alpha(cf_{\theta_0}, f_\theta), \quad (24)$$

where the approximation holds provided that $(1-c) \int f_\theta^\alpha h$ is small enough (Kanamori and Fujisawa, 2015). Unlike $D_\gamma(g, f_\theta)$ in (10), where the mixing proportion c appears outside

$D_\gamma(f_{\theta_0}, f_\theta)$, here the mixing proportion c appears inside $D_\alpha(cf_{\theta_0}, f_\theta)$. This effect leads to less robustness of α -logistic compared with γ -logistic.

The difference between two methods can be further clarified via comparing the misclassification rate of the prediction rule $y = I(\hat{\beta}_\bullet^\top x > 0)$, where $\hat{\beta}_\bullet$ can stand for either $\hat{\beta}_\gamma$ or $\hat{\beta}_\alpha$. Croux, Haesbroeck and Joossens (2008) showed that the robustness of misclassification rate is characterized by its second order influence function $\text{IF2}_{\hat{\beta}_\bullet}(x, y)$. The second order influence function for a functional $T(F)$ of the distribution F at z is $\frac{\partial^2}{\partial \varepsilon^2} T\{(1 - \varepsilon)F + \varepsilon\delta_z\}|_{\varepsilon=0}$, where δ_z is the Dirac measure at z . For the case of $p = 2$ with $X_1 = 1$, $\beta_0 = (\beta_{01}, \beta_{02})^\top$, and $\hat{\beta}_\bullet = (\hat{\beta}_{\bullet 1}, \hat{\beta}_{\bullet 2})^\top$, one has $\text{IF2}_{\hat{\beta}_\bullet}(x, y) \propto \{\beta_{01} \text{IF}_{\hat{\beta}_{\bullet 2}}(x, y) - \beta_{02} \text{IF}_{\hat{\beta}_{\bullet 1}}(x, y)\}^2$, which is plotted in Figure 2 with various $\gamma = \alpha$ values, where $\text{IF}_{\hat{\beta}_{\bullet j}}(x, y)$ is the influence function of $\hat{\beta}_{\bullet j}$, $j = 1, 2$. When $\gamma = 0$, both methods reduce to the non-robust MLE, and an unbounded $\text{IF2}_{\hat{\beta}_\bullet}(x, y)$ is detected. Note that $\text{IF2}_{\hat{\beta}_\bullet}(x, y)$ has larger value at non-matched (x, y) value, which reflects the influence of outliers. We also detect that $\text{IF2}_{\hat{\beta}_\bullet}(x, 0) > 0$ around $x = -1$. This is reasonable since $P(Y_0 = 1) = 2P(Y_0 = 0)$ in our setting, which gives more samples with $Y_0 = 1$. As a result, a data point from the 0-group is expected to be more influential than that from the 1-group. When $\gamma = 0.5$, $\text{IF2}_{\hat{\beta}_\bullet}(x, y)$ at non-matched (x, y) are largely reduced, indicating the robustness of γ -logistic and α -logistic to mislabeling. The difference between two methods becomes clear when $\gamma \geq 1.5$, where $\text{IF2}_{\hat{\beta}_\alpha}(x, y) > 0$ for a wide range of x , while $\text{IF2}_{\hat{\beta}_\gamma}(x, y) > 0$ at limited region of x only. That is, γ -logistic becomes more and more resistant to mislabeling as γ increases. The robustness of γ -logistic, as mentioned in Section 2.1, comes from the locality nature of γ -divergence. It also implies that, when γ is large, the performance of γ -logistic is mainly determined by data points near the decision boundary $\beta_0^\top x = 0$ ($x = -\ln 2$ in this case). This explains the observation at $\gamma \geq 1.5$ that $\text{IF2}_{\hat{\beta}_\gamma}(x, y)$ can have larger value than $\text{IF2}_{\hat{\beta}_\alpha}(x, y)$, especially when $y = 0$ (i.e., the 0-group with fewer samples).

Remark 7. *There exist robustified logistic regression methods other than the constant-mislabel logistic, ξ -logistic, and α -logistic. A majority of them have a robust estimating equation of the form $\frac{1}{n} \sum_{i=1}^n [w_i(\beta) \{Y_i - \pi(X_i; \beta)\} - b(X_i; \beta)] X_i = 0$, where the weight $w_i(\beta)$ can depend on (X_i, Y_i) . The bias correction term $b(X_i; \beta)$ is used to ensure Fisher consistency in the presence of $w_i(\beta)$. See Bianco and Yohai (1996), Carroll and Pederson (1993), Stefanski, Carroll, and Ruppert (1986), Künsch, Stefanski, and Carroll (1989) among others for different choices of $w_i(\beta)$. Note that γ -logistic does not belong to this class, since it uses $\{Y_i - \pi(X_i; (\gamma + 1)\beta)\}$ for bias correction.*

4 Numerical Studies

4.1 Simulation settings

We use the Pima data (see Section 5 for details) to conduct simulation studies. In each simulation run, $n = 500$ covariate vectors $X_0 \in \mathbb{R}^8$ are randomly sampled from the Pima data (after component-wise standardization) and $X = (X_0^\top, 1)^\top$. Given X , the response variable Y is generated from (3) with the following settings of mislabel probabilities: (S1) $\eta_0(x) = u_0$ and $\eta_1(x) = u_1$; (S2) $\eta_0(x) = \eta_1(x) = u_0 + (u_1 - u_0) \frac{\exp(\beta_0^\top x)}{1 + \exp(\beta_0^\top x)}$; (S3) $\eta_j(x) = u_0 + (u_1 - u_0) \frac{\exp(b_j^\top x)}{1 + \exp(b_j^\top x)}$, where each element of $b_j \in \mathbb{R}^9$, $j = 0, 1$, is generated from $N(0, 2^2)$ for each simulation; and (S4) $\eta_0(x) = u_0 + (u_1 - u_0)I(|X_1 - a| < 3, |X_3 + a| < 3)$ and $\eta_1(x) = u_0 + (u_1 - u_0)I(|X_1 + a| < 3, |X_2 + a| < 3)$, where $a \sim N(2, 0.3^2)$ for each simulation. Setting (S1) considers Y_0 -dependent mislabeling. Setting (S2) considers X -dependent mislabeling, where mislabeling is more likely to occur for subjects with higher success probability. Settings (S3)-(S4) consider (Y_0, X) -dependent mislabeling. In (S3) $\eta_j(x)$'s depend on random linear combinations of X . In (S4) mislabeling is more likely to occur for (X_1, X_3) around $(a, -a)$ when $Y_0 = 0$, and also more likely to occur for (X_1, X_2)

around $(-a, -a)$ when $Y_0 = 1$. We set $u_0 = 0.05$ and $u_1 \geq 0.05$ such that in all settings, $u_1 = u_0$ indicates that the constant-mislabel logistic holds, while $u_1 > u_0$ indicates a deviation from it.

Two types of γ selection are implemented. One is based on the data-adaptive method in Remark 4 (denoted by γ -logistic). The other is based on an independent uncontaminated data $\{(X_i^*, Y_{0i}^*)\}_{i=1}^n$ that selects γ by maximizing the likelihood $\prod_{i=1}^n \pi(X_i^*; \hat{\beta}_\gamma)^{Y_{0i}^*} \{1 - \pi(X_i^*; \hat{\beta}_\gamma)\}^{1-Y_{0i}^*}$ (denoted by γ^* -logistic). Of course Y_{0i}^* 's are not observed, and γ^* only represents an ideal γ value for comparison purpose. In addition to γ -logistic and γ^* -logistic, we also implement the conventional logistic regression (denoted as logistic), constant-mislabel logistic, ξ -logistic, and α -logistic (where α is optimally tuned as γ^* -logistic does, and it is denoted by α^* -logistic). Simulation results are reported with 500 replicates.

4.2 Simulation results

We first evaluate the performances of $(\hat{\beta}_\gamma, \hat{\Sigma}_\gamma)$. Simulation results for $\gamma = 2$ under (S1)-(S2) with $\beta_0 = (0, 1, -1, 1, \mathbf{0}_{p-3}^\top)^\top$ and $u_1 = 0.1$ are placed in Table 1, which reports the means of $\hat{\beta}_\gamma$ (Mean), the standard deviations of $\hat{\beta}_\gamma$ (SD), and the means of the diagonal elements of $\hat{\Sigma}_\gamma$ (SE) over 500 replicates. One can see that $\hat{\beta}_\gamma$ targets β_0 with only small bias under both mislabeling mechanisms (S1)-(S2). Moreover, SE are found to be close to SD, which shows the validity of the proposed sandwich-type estimator $\hat{\Sigma}_\gamma$.

We next compare the performance of γ -logistic with other methods. The values of γ and α are selected over $[0.5, 2.5]$ with $\gamma_0 = 0.1$. In this simulation, each element of β_0 is generated from $N(0, 2^2)$ for each replicate. Figure 3 reports the classification accuracy (CA) from applying the prediction rule $y = I(\hat{\beta}_\gamma^\top x > 0)$ to an independent clean data (Y_0, X) with size n , where the x -axis represents the corresponding mislabel rate $\tau = P(Y \neq Y_0)$ under $u_1 \in \{0.05, 0.1, \dots, 0.5\}$. We also report in Table 2 the means of the selected γ and γ^*

values of γ -logistic and γ^* -logistic. Observe that the robustified logistic methods (γ -logistic, α -logistic, constant-mislabel logistic) dominate the conventional logistic under (S1)-(S4), but not the ξ -logistic. Recall that ξ -logistic assumes that mislabeling tends to occur for subjects lying near the decision boundary $\beta_0^\top x = 0$. This assumption is not satisfied in (S1)-(S4). As a result, ξ -logistic can perform even worse than the conventional logistic regression under (S3)-(S4), especially for the case of severe mislabeling (i.e., large τ). It conveys an important message that, while incorporating a correct mislabeling mechanism into the estimation method can be beneficial, the correctness of model specifications for $\eta_j(x)$'s is critical to the analysis result. Misspecifying $\eta_j(x)$'s can sometimes lead to worse result. However, γ -logistic, which avoids modeling $\eta_j(x)$'s, is able to adapt to various mislabeling mechanisms and can be less affected by model misspecification.

We now compare γ -logistic with constant-mislabel logistic and α -logistic. For small τ , the constant-mislabel assumption approximately holds and constant-mislabel logistic produces the highest CA values as expected, while γ -logistic has comparable performances. For large τ , the mislabeling mechanism becomes complicated, which adversely affects the performances of constant-mislabel logistic. In this case, γ -logistic produces the highest CA values under (S1)-(S4). It is also found that γ -logistic outperforms α^* -logistic, even α^* -logistic selects α optimally. Recall the comparison discussions of robustness for γ -logistic and α -logistic in Section 3.3. Our simulation results confirm the superiority of γ -logistic in dealing with various mislabeling mechanisms. Finally, comparing γ -logistic with the optimal γ^* -logistic, the loss of γ -logistic from using the data-adaptive γ is not large, indicating the applicability of the proposed data-adaptive selection criterion of γ .

5 The Pima Data Analysis

The Pima data (available from the *UCI machine learning repository*) contains females of Pima Indian heritage, each with an indicator of diabetes status (Y) and 8 covariates (standardized to have mean 0 and variance 1), including the pregnant times (X_1), glucose concentration (X_2), blood pressure (X_3), triceps skin fold thickness (X_4), serum insulin (X_5), BMI (X_6), diabetes pedigree function (X_7), and age (X_8). We set $X_9 = 1$ to include the intercept term. Detailed description of the data can be found in Smith *et al.* (1988). Medical data can more easily suffer the problem of mislabeling, and we aim to use the robust γ -logistic to investigate the effects of these covariates on the diabetes status.

Figure 4 (a) provides the estimates $\hat{\beta}_\gamma$ from γ -logistic together with the 95% confidence intervals. Figure 4 (b) provides the estimated success probabilities $\pi(X_i; \hat{\beta}_\gamma)$'s for two groups. The analysis results from the conventional logistic regressions are also placed in Figure 4 (c)-(d) for comparison. In general, γ -logistic tends to produce wider confidence intervals than conventional logistic. This is expected since the robustness of γ -logistic comes at the cost of being less efficient than MLE. Both analysis results show that (X_1, X_2, X_6, X_7) are critical (significant or nearly significant) factors to the diabetes status. Interestingly, γ -logistic further demonstrates that (X_3, X_5) are significant factors (as the corresponding confidence intervals do not contain 0), and X_4 is nearly significant. Considering the robustness of γ -logistic, this difference would mainly result from treating some instances as outliers, by assigning them less weights during model fitting. In particular, we obtain more precise estimates for the effects of blood pressure (X_3), triceps skin fold thickness (X_4), and serum insulin (X_5) when possible mislabeled subjects have been weighed down.

From the results of γ -logistic, instances with $PV_i < 0.01$ are marked with “+” in Figure 4 (b). These instances are candidates of mislabeled subjects. To further investigate the driven factors of mislabeling, we define the mislabeling status $\delta_i = I(PV_i < 0.01)$,

and then estimate the true response by $\widehat{Y}_{0i} = Y_i(1 - \delta_i) + (1 - Y_i)\delta_i$, i.e., subjects with $\delta_i = 1$ are flipped for label correction. We then fit the conventional logistic regression to $(\delta_i, X_i)|\widehat{Y}_{0i} = j$ to obtain the regression coefficient b_j for $j = 0, 1$. Note that b_j quantifies how X affects the chance of being mislabeled within the group of $Y_0 = j$. The AUC values from $(\delta_i, b_0^\top X_i)|\widehat{Y}_{0i} = 0$ is 0.713, while it is 0.925 from $(\delta_i, b_1^\top X_i)|\widehat{Y}_{0i} = 1$. It indicates that X is influential to the mislabel probability $\eta_1(x)$, while the mislabel probability $\eta_0(x)$ tends to be constant for subjects without diabetes. Moreover, the result of $b_1 = (0.258, 0.322, 0.491, -0.837, 0.303, 1.106, 1.061, 0.512, -6.989)$ suggests (X_6, X_7) (with p-values smaller than 0.05) to be possible driven factors of mislabeling for diabetes patients.

6 Discussion

In this work we only consider the case of mislabeling in the response Y , while X is assumed to be uncontaminated. In the presence of leverage points of X that are influential to the final estimates, γ -logistic can be modified to mitigate the effects of outlying X_i by using a weighting scheme $w_{\gamma,i}(\beta)q(X_i)$ in the estimating equation (16). For example, Croux, Haesbroeck and Joossens (2008) suggested $q(x) = I\{(x - \mu_X)^\top \Sigma_X^{-1}(x - \mu_X) \leq a\}$ for some user-defined constant a , where μ_X and Σ_X are some robust estimates of $E(X)$ and $\text{cov}(X)$. Since $q(X_i)$ does not depend on Y_i , Theorem 5 still holds for the modified γ -logistic by replacing H_γ and U_γ with $E[q(X_i)\|f(\cdot|X_i; \beta_0)\|_{\gamma+1}\nu(X_i; (\gamma+1)\beta_0)X_iX_i^\top]$ and $E[q^2(X_i)w_{\gamma,i}^2(\beta_0)\{Y_i - \pi(X_i; (\gamma+1)\beta_0)\}^2X_iX_i^\top]$, respectively. It is of interest to investigate the choice and effect of $q(\cdot)$ on γ -logistic in a future study.

For the purpose of robustness, Ghosh and Basu (2016) developed a robust GLM using the density power divergence, which includes α -logistic as a special case. We have shown that γ -logistic outperforms α -logistic under severe mislabeling in numerical studies. The developed methodology (11)-(13) can be extended to robust GLM, including multi-class

Y (see Supplementary Materials for a brief illustration), count Y , and continuous Y . Although the idea is straightforward, further efforts are required to develop the validity of the approximation (10), the asymptotic properties, and the implementation algorithms. It is also of interest to compare the differences between the robust GLM using γ -divergence and the robust GLM of Ghosh and Basu (2016) using density power divergence.

References

- Bianco, A. M. and Yohai, V. J. (1996). Robust estimation in the logistic regression model. *In Robust statistics, data analysis, and computer intensive methods* (pp. 17-34). Springer New York.
- Carroll, R. J. and Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society, Series B*, 55, 693-706.
- Copas, J. B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society, Series B*, 50, 225-265.
- Croux, C., Haesbroeck, G., and Joossens, K. (2008). Logistic discrimination using robust estimators: an influence function approach. *Can. J. Stat.*, 36, 157-174.
- Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99, 2053-2081.
- Ghosh, A. and Basu, A. (2016). Robust estimation in generalized linear models: the density power divergence approach. *Test*, 25, 269-290.
- Hayashi, K. (2012). A boosting method with asymmetric mislabeling probabilities which depend on covariates. *Computational Statistics*, 27, 348-356.

- Jones, M. C., Hjort, N. L., Harris, I. R. and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika*, 88, 865-873.
- Kanamori, T. and Fujisawa, H. (2015). Robust estimation under heavy contamination using unnormalized models. *Biometrika*, 102, 559-572.
- Komori, O., Eguchi, S., Ikeda, S., Okamura, H., S., Ichinokawa, M., and Nakayama, S. (2016). An asymmetric logistic regression model for ecological data. *Methods in Ecology and Evolution*, 7, 249-260.
- Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, 84, 460-466.
- Mollah, M. N. H., Eguchi, S., and Minami, M. (2007). Robust prewhitening for ICA by minimizing β -divergence and its application to FastICA, *Neural Process Lett.*, 25, 91-110.
- Stefanski, L. A., Carroll, R. J., and Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika*, 73, 413-424.
- Smith, J., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Symposium on Computer Applications and Medical Care*, 9, 261-265.
- Takenouchi, T. and Eguchi, S. (2004). Robustifying AdaBoost by adding the naive error rate. *Neural Computation*, 16, 767-787.
- Wainer, H., Bradlow, E. T., and Wang, X. (2007). *Testlet Response Theory and Its Applications*. Cambridge University Press, New York.

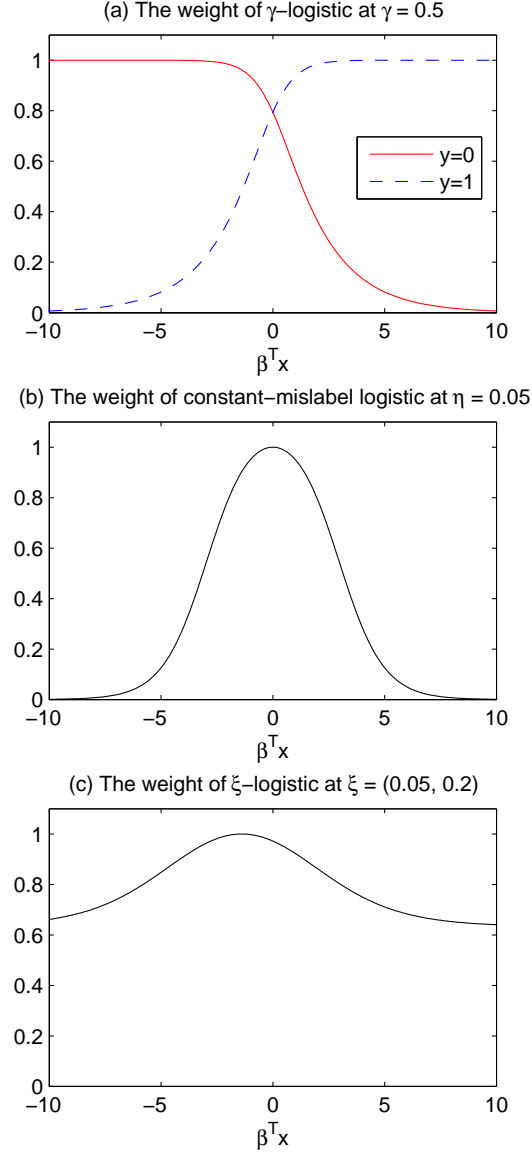


Figure 1: The weight functions (scaled to have a maximum value 1) of (a) γ -logistic with $\gamma = 0.5$, (b) constant-mislabel logistic with $\eta = 0.05$, and (c) ξ -logistic with $\xi = (0.05, 0.2)$.

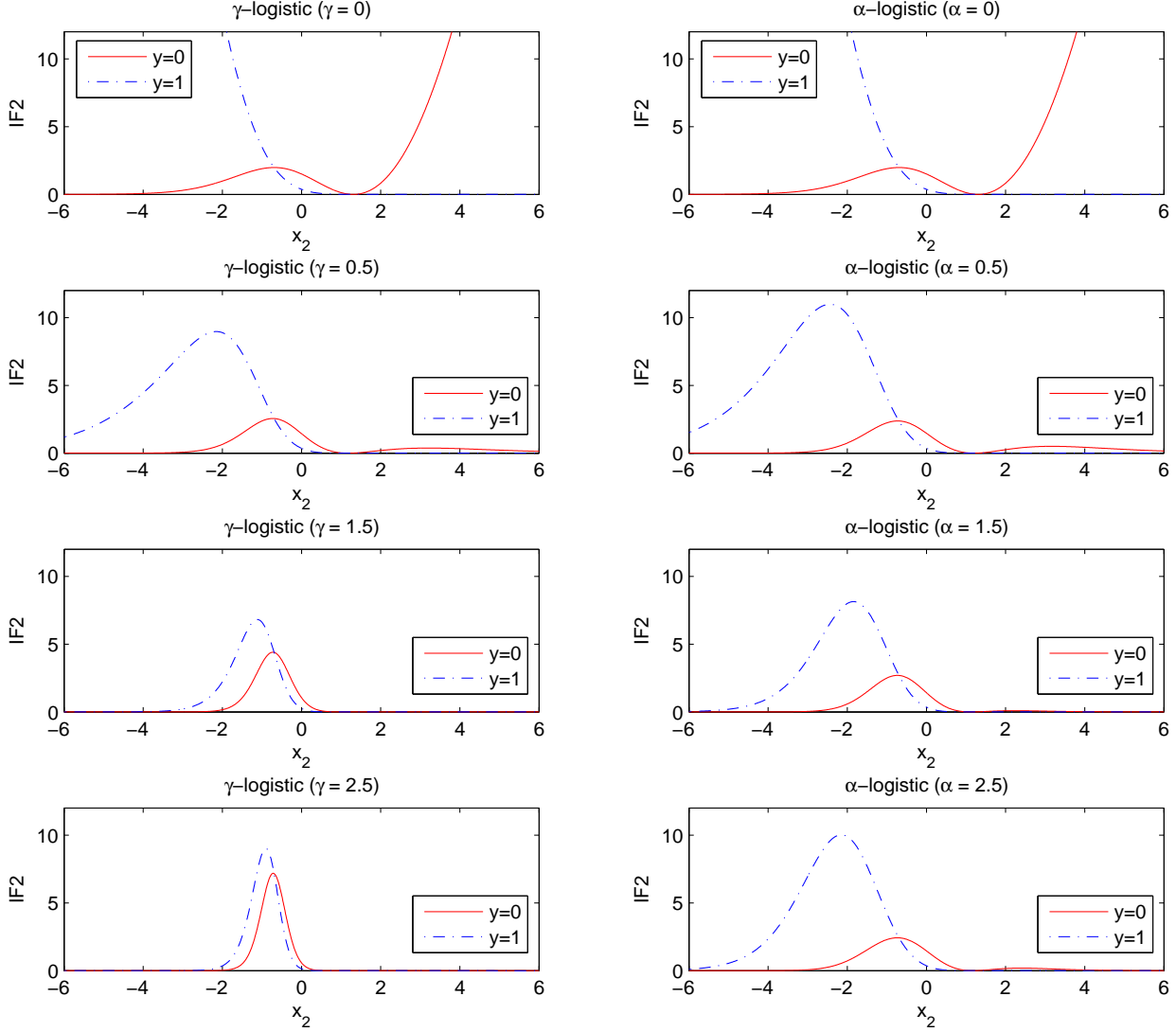


Figure 2: The second-order influence functions of misclassification rate of γ -logistic (the left panel) and α -logistic (the right panel) at $\gamma = \alpha \in \{0, 0.5, 1.5, 2.5\}$, where the real line is for the case of $y = 0$, and the dash-dotted line is for the case of $y = 1$. The plots are obtained under the setting of $p = 2$, where $X_1 = 1$ is the intercept term, $X_2|Y_0 = 0 \sim N(-0.5, 1)$, $X_2|Y_0 = 1 \sim N(0.5, 1)$ and $P(Y_0 = 1) = 2P(Y_0 = 0)$. It gives $\beta_0 = (\ln 2, 1)^\top$.

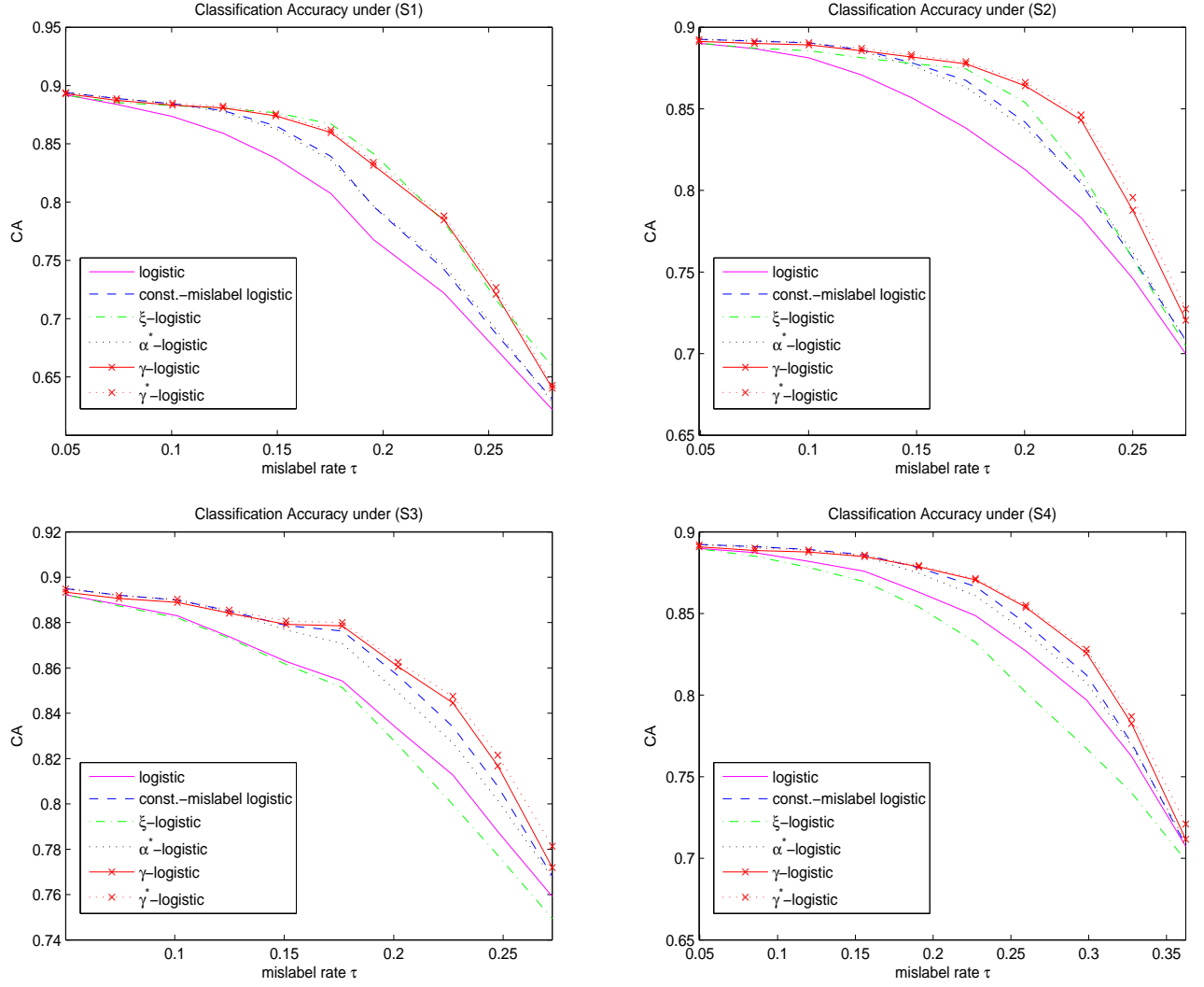


Figure 3: Simulation results of the classification accuracy (CA) under (S1)-(S4) with different values of u_1 , where the x -axis represents the corresponding mislabel rate $\tau = P(Y \neq Y_0)$.

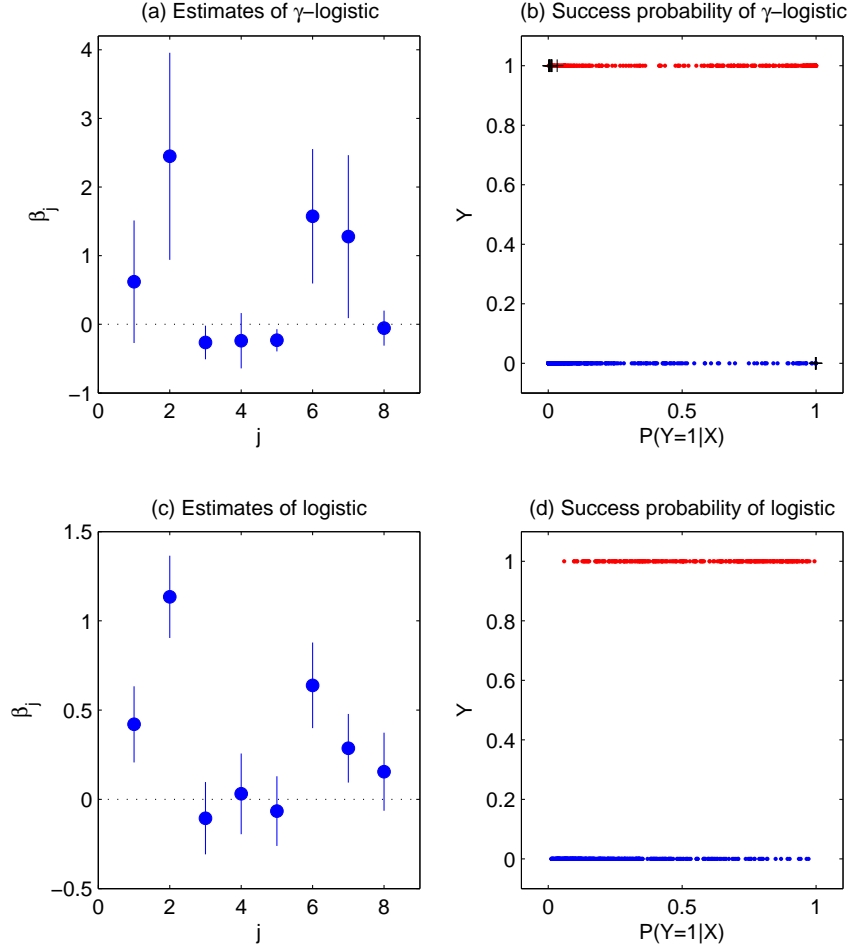


Figure 4: (a): The regression coefficients $\hat{\beta}_\gamma$ from γ -logistic, where the vertical lines represent the 95% confidence intervals. (b) The success probabilities $\pi(X_i; \hat{\beta}_\gamma)$'s for two groups $Y_i = 1$ and $Y_i = 0$ from γ -logistic. Subjects with $PV_i < 0.01$ are marked with “+”. The analysis results from conventional logistic are placed in (c)-(d).

Table 1: The means of $\widehat{\beta}_\gamma$ (Mean), the standard deviations of $\widehat{\beta}_\gamma$ (SD), and the means of the diagonal elements of $\widehat{\Sigma}_\gamma$ (SE) from γ -logistic under settings (S1)-(S2).

	(S1)				(S2)			
	True	Mean	SD	SE	True	Mean	SD	SE
β_{01}	0.000	-0.126	0.171	0.171	0.000	-0.014	0.168	0.169
β_{02}	1.000	1.009	0.308	0.323	1.000	0.999	0.307	0.326
β_{03}	-1.000	-0.999	0.312	0.316	-1.000	-0.995	0.296	0.315
β_{04}	1.000	1.014	0.307	0.320	1.000	0.984	0.281	0.317
β_{05}	0.000	-0.025	0.208	0.206	0.000	0.011	0.201	0.206
β_{06}	0.000	-0.033	0.224	0.192	0.000	-0.006	0.217	0.197
β_{07}	0.000	0.013	0.209	0.210	0.000	-0.001	0.223	0.216
β_{08}	0.000	0.008	0.185	0.176	0.000	-0.018	0.190	0.184
β_{09}	0.000	0.004	0.211	0.203	0.000	0.014	0.220	0.208

Table 2: The means of the selected γ and γ^* values at different u_1 under (S1)-(S4).

u_1		0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
(S1)	γ^*	0.97	1.27	1.60	1.82	2.01	2.12	2.10	2.02	1.72	1.20
	γ	1.26	1.60	1.84	2.04	2.20	2.32	2.30	2.21	1.92	1.76
(S2)	γ^*	0.99	1.28	1.60	1.82	2.02	2.16	2.22	2.21	1.87	1.36
	γ	1.30	1.59	1.87	2.05	2.20	2.30	2.37	2.28	1.95	1.64
(S3)	γ^*	1.02	1.26	1.56	1.79	2.00	2.14	2.17	2.22	2.07	1.77
	γ	1.35	1.57	1.84	2.08	2.20	2.31	2.37	2.40	2.38	2.34
(S4)	γ^*	0.96	1.38	1.78	2.04	2.29	2.39	2.39	2.32	2.04	1.55
	γ	1.29	1.74	2.03	2.24	2.39	2.46	2.45	2.35	2.11	1.87